International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

GPD Threshold Estimation Using Measure of Surprise

Abraham Manurung[a]*, Aji Hamim Wigena[b], Anik Djuraidah[c]

[a,b,c]*Department of Statistics, Faculty of Mathematics and Natural Science, Bogor Agricultural University, Bogor, Indonesia*

[a]*Email: abemadison48@gmail.com*

[b]*Email: ajiwigena@ymail.com* , [c]*Email: anikdjuraidah@gmail.com*

Abstract

Threshold is used to estimate parameters of Generalized Pareto distribution to estimate return value. This return value shows the extreme value in the period of time. Threshold can be estimated using Mean Residual Life Plot, Threshold Stability Plot, or the upper 10% rule but this estimation is usually subjective. An alternative method is measure of surprise based on Bayes method and Monte Carlo Markov Chain technique. This paper aims to estimate a GPD threshold based on simulation data and to apply measure of surprise method to rainfall data in the period of 1981-2012 in Bogor, Indonesia. The simulation result showed that the predicted threshold is exactly the same as the true threshold. The result of application to rainfall data showed the threshold was approximately 210 mm.

Keywords: Bayes; GPD; measure of surprise; posterior predictive distribution; rainfall; threshold estimation.

1. Introduction

Extreme Value Theory (EVT) is one of the branches in statistical science that is used to measure probabilities in an extreme event, both large or small extreme. The application is implemented in various fields such as risk assessment on financial markets, wind engineering, food science, biomedical data processing, and assessment of meteorological change [1]. This large or small unusual behavior happens at the tail of a probability distribution. According to Vicari [2], EVT is a statistical discipline that models and studies the tail of distributions. One distribution which is often used for extreme quantity estimation is a Generalized Pareto distribution (GPD).

---

* Corresponding author.

This distribution was first introduced by [3] and developed comprehensively by Davidson and Smith [4]. Estimating GPD parameter starts with determining a boundary between the extreme and non-extreme values in a dataset. This boundary will be called the threshold. The group of data which is categorized extreme based on the threshold will be separated from the original dataset to be a new dataset for the extreme value analysis. Hence, GPD parameter estimation depends on the threshold. Behrens [5] presents the dilemma that if a high threshold is chosen then only a few observations can be used to estimate the GPD parameter, thus the variance of the estimates will increase. However, if the selected threshold is too low then a bias could occur because a non-extreme observation is categorized as extreme.

The method to select a threshold is called Peak Over Threshold (POT). In the last three decades, POT has been developed in many approaches with their advantages and disadvantages. Scarrott and MacDonald [6] emphasize the importance of selecting a good threshold which could balance the bias between having too few or too many observations. They made a comprehensive review on extreme value threshold estimation history until 2012. The method that is widely used is a graphical diagnostic. Davidson and Smith [4] introduced Mean Residual Life Plot (MRLP). Threshold Stability Plot was popularized by Coles [1]. A key drawback of these methods is that it needs an assessment expertise and the graph interpretation is usually subjective. Another approach by [7] is called the upper 10% rule which proposes to use the 9^{th} quantile as the threshold. This method is theoretically inappropriate and depends highly on the number of observations.

Based on the drawback of the aforementioned methods, other alternatives had been developed including the measure of surprise method. The idea of surprise was first proposed by Weaver [8], then consolidated by Box [9], Bayarri and Berger [10], and Meng [11].

The measure of surprise concept in threshold estimation for extreme data was applied by Vicarri [2] and Lee and her colleagues [12]. In summary, threshold estimation using measure of surprise is one of the latest methods and is a method which is fully Bayesian. The aim of this research is to apply threshold estimation on simulation data and real data (Bogor rainfall) using measure of surprise and to compare this method with the widely used MRLP.

2. Literature Reviews

2.1 Generalized Pareto Distribution

The GPD is a suitable distribution to analyze extreme data by the POT approach. Coles [1] stated if $X_1, \dots, X_n \in \mathbb{R}$ is a sequence of independent and identical random variables, then the asymptotic distribution of the exceedances, $Y = X - u \mid X > u$ of some high threshold u is a GPD given by:

$$F(y) = 1 - \left[1 + \xi \frac{(y-u)}{\sigma}\right]_+^{-1/\xi} \quad (1)$$

with $[a]_+ = \max\{0,a\}$ where $\sigma > 0$ is the scale parameter, and $-\infty < \xi < \infty$ is the shape parameter. The result of GPD parameter estimation is used to determine the GPD return value.

2.2 Measure of Surprise

Lee and her colleagues [13] stated that measure of surprise is used to quantify the degree of incompatibility of observed data and the given model. Once a model $H_0: x_{obs} \sim f(x|\theta)$ is formulated and x_{obs} is observed, Bayarri dan Berger [10] asked how surprising the observation is. In the threshold estimation context, $f(x|\theta)$ corresponds with GPD. The incompatibility degree is measured by a surprise value, namely the p-value:

$$p = \Pr_f(T(x) \geq T(x_{obs})) \quad (2)$$

$T(X)$ is a certain test statistic to investigate the incompatibility between the observed data and the model (GPD). The test statistic used in this GPD threshold estimation case is a reciprocal likelihood function: $T(x) = 1/f(x|\theta)$. In Bayesian methods, posterior distribution is calculated by doing integration of the following: Posterior $\propto$ Prior x Likelihood

$$\pi(\theta|\boldsymbol{x}) \propto p(\theta)L(\theta|\boldsymbol{x}) \quad (3)$$

Then, a *posterior predictive distribution* which contains $f(x|\theta)$ was first proposed by Guttman [13] is given as:

$$m(x|x_{obs}) = \int f(x|\theta)\pi(\theta|x_{obs})\, d\theta \quad (4)$$

where $\pi(\theta|x_{obs}) \propto f(x_{obs}|\theta)\pi(\theta)$ is a posterior distribution for θ if x_{obs} is known. Having both test statistics, where posterior predictive distribution $T(x)$ comes from GPD theoretical distribution (by simulation) and posterior predictive distribution $T(x_{obs})$ from observed data, will lead to the posterior predictive p-value:

$$p_{post} = \Pr_{mo}(T(x) \geq T(x_{obs})) \quad (5)$$

This posterior predictive p-value (p_{post}) is called the surprise. A high *p-value* which is caused from a large value of $T(x)$ shows a high suprise. This means that we would be surprise if the observed data, x_{obs}, is compatible with the given model H_0 (GDP in this case). Therefore, we can safely interpret p-values which are close to 1 as an indicator of x_{obs} and H_0 incompatibility. Meng [11] stated that the posterior predictive p-value that is expected of x_{obs} and H_0 compatibility is 0.5. In this research, we used a reciprocal likelihood function which reverses the p-value interpretation: where 0 is considered as a surprise. The GPD threshold selection concludes when the surprise converges at 0.5. This point of convergence is considered as the minimum threshold.

3. Data and Methodology

3.1 Data and Tools

There two different datasets which are used in this research: simulation data and rainfall data from the Meteorological, Climatological, and Geophysical Agency (MCGA) of Indonesia. The description of both data is as follows:

1. Simulation dataset is generated from a mixture distribution X_{obs} ~ 0.7 Gamma (α=10, β=8, u=100) + 0.3 GPD (σ=40, ξ=0.1), n=2400. Gamma is truncated at u=100 as a known threshold.
2. MCGA dataset is 10-day rainfall data for 32 years (1981-2012) in Bogor. The total observations for 32 years with 36 observations per year are 1152 observations. Rainfall unit is millimeter (mm).

This research uses R 3.4.0 software for data analysis. The packages to produce MRLP is 'ismev', to generate random data of GPD is 'SpatialExtremes', and to generate the graphs is 'ggplot2'. Measure of surprise is calculated without using R package, instead of developing R code by applying Bayesian methods and MCMC Metropolis-Hastings algorithm.

3.2 Methodology

The steps of data analysis are as follows:

1. Explore simulation and rainfall dataset with histogram and density plot. Simulation and rainfall data will be address as the observed data (x_{obs}).

2. Define a series of candidate thresholds (U_i) where $U_1 < \cdots < U_s$

   a. The threshold candidates (U_i) for simulation dataset are 0 - 200 by 20 interval (s=11).

   b. The threshold candidates (U_i) for rainfall dataset are 0 - 280 by 10 interval (s=29).

3. For every threshold candidate (U_i):

   i. Determining the number of Markov chain on the MCMC Metropolis-Hastings algorithm: 10.000 chains for each dataset.
   ii. On every chain, based on Bayesian method, estimate GPD parameters $\boldsymbol{\theta} = (\sigma, \xi)$ by calculating the posterior distribution $\pi(\boldsymbol{\theta}|x_{obs}(U_i))$ where $x_{obs}(U_i) = \{x_{obs} : x_{obs} > U_i\}$ is the subset of elements of x_{obs} that exceed U_i [13].

- The prior used for the calculation is a Jeffrey's prior. Castellanos and Cabras [14] state Prior Jeffrey for GPD distribution as:

$$\pi(\boldsymbol{\theta}) \propto \sigma^{-1}(1+\xi)^{-1}(1+2\xi)^{-1/2} \quad (6)$$

- The log-likelihood function, which is the product of the $f(x_i|\boldsymbol{\theta})$, can be approached by calculating the mean of the $f(x_i|\boldsymbol{\theta})$:

$$f(\boldsymbol{x}|\boldsymbol{\theta}) \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \quad (7)$$

- The Bayesian integration process is calculated in a log scale which forms the final formula to obtain the posterior distribution as:

$$\log(\pi(\boldsymbol{\theta}|x_{obs})) = f(\boldsymbol{x}|\boldsymbol{\theta}) - \log(\sigma) - \log(1+\xi) - 0.5\log(1+2\xi) \quad (8)$$

The posterior distribution using MCMC Metropolis-Hastings algorithm will produce 10.000 chains of $\boldsymbol{\theta}$ values with the first chain (initial parameter value) as:

a. Simulation dataset initial parameter values : $\xi_0 = 0.3$; $\sigma_0 = 38$

b. Rainfall dataset initial parameter values : $\xi_0 = 0.02$; $\sigma_0 = 40$

On the MCMC Metropolis-Hastings algorithm, the proposal density for σ is a log-normal density distribution and for ξ is a normal density distribution.

iii. Draw random sample parameter from the produced chain: 1000 for simulation dataset and 5000 for rainfall dataset. The samples are taken from the 2501^{th} – 10000^{th} chain (the first 2500 is a burn-in period chain).

iv. Calculate $m(x)$ and $m(x_{obs})$ using the drawn random sample parameter (based on equation 4).

v. Calculate posterior predictive p-value which is the proportion $m(x) \leq m(x_{obs})$.

4. Repeat step 3 for 30 times (for both observed data)

5. Construct a boxplot between the threshold and the p-values and determine the most suitable threshold.

6. Compare the measure of surprise method to the MRLP.

4. Result and Discussion

4.1 Simulation Case

First of all, we would like to see how to measure of surprise can be used in simulated data where the true threshold is known (u = 100). The results that will be shown is the histogram and density plot, a measure of surprise diagram, and MRLP.

The generated simulation data histogram and density plot which is a mixture distribution are in Figure 1. The total number of the observation data is 2400 with 1693 observation under the threshold and 707 above. The red dashed line shows the true threshold. The transition from truncated Gamma distribution to GPD is very subtle which makes it hard to determine the threshold if we only use a mere data graph exploration.

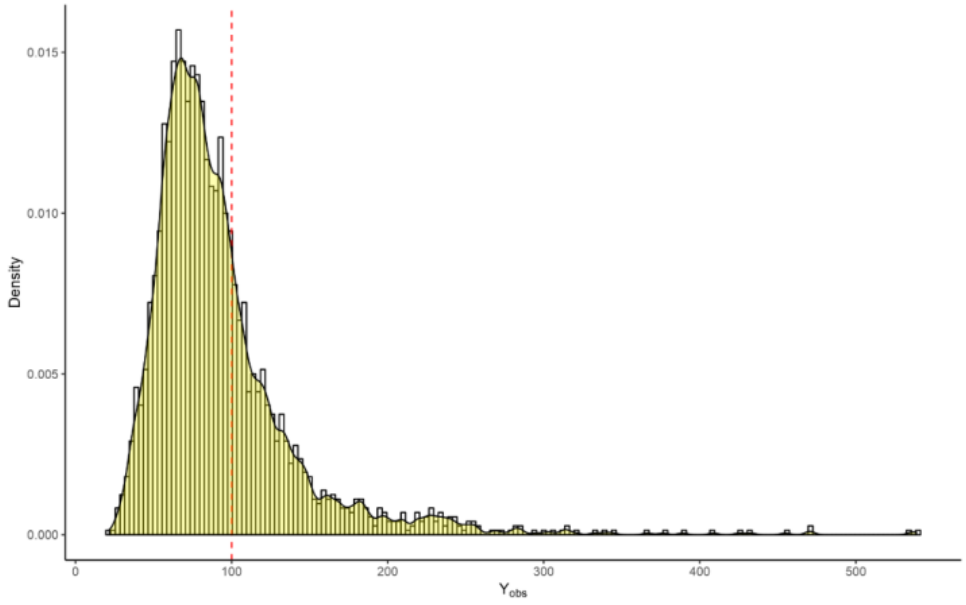

Figure 1: Histogram and Density Plot of Simulation Data

Measure of surprise diagram for simulation case is in Figure 2. There are 11 threshold candidates with 30 repetition calculation of the surprise (posterior predictive p-value) on every candidate. The surprise is between 0 and 1 where surprise near 0 (because of reciprocal likelihood function) shows that we are surprised by the fact that the distribution on that point is compatible with GPD. Surprises on 0, 20, and 40 threshold candidate is 0 which means we are highly surprised. This indicates that the 0, 20, and 40 candidate is not compatible with GPD. Surprise start to increase at 60 (p=0.14) and 80 (p=0.42) threshold candidate which means the distribution at those points is starting to be more compatible with GPD. Surprises on 100 until 200 is stable around 0.5 which means all of the points on this range is compatible with GPD. Finally, the measure of surprise diagram in Figure 2 shows that the threshold on the mixture distribution is the same with the true threshold which is 100.

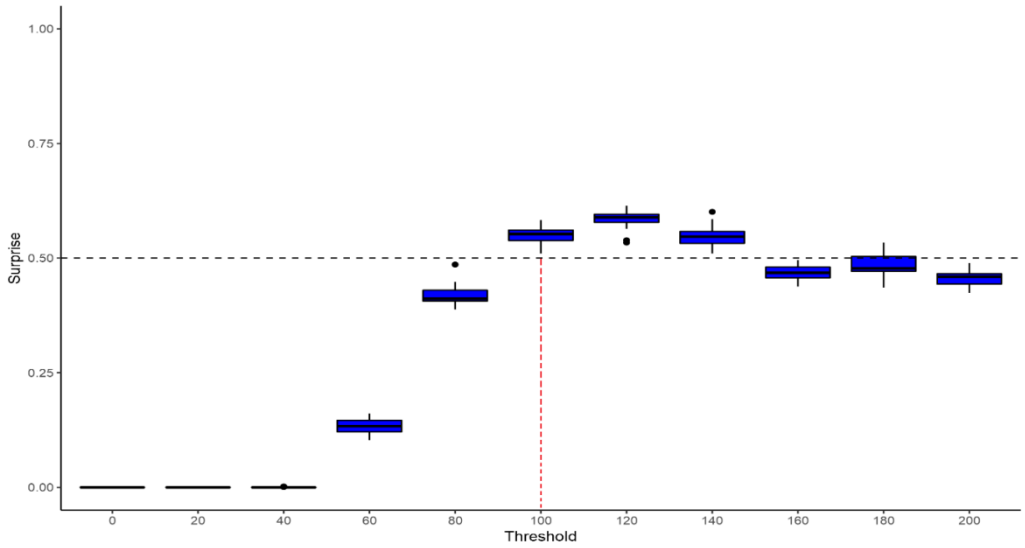

Figure 2: Measure of Surprise Diagram for Simulation Data

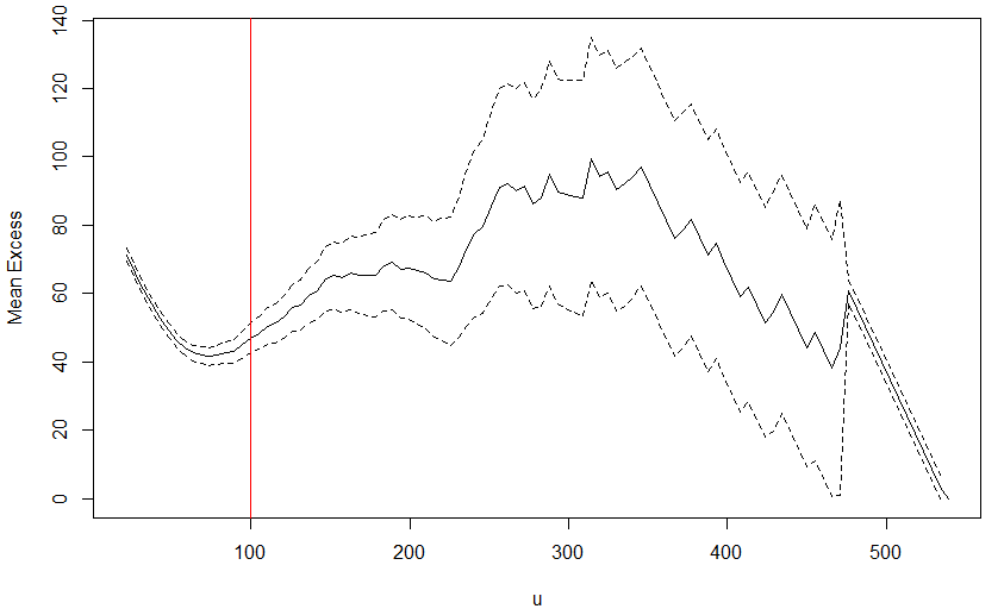

Figure 3: Mean Residual Life Plot for Simulation Data

The next step is comparing measure of surprise diagram with MRLP as the most used classical graph method. MRLP is interpreted by seeing the line between threshold and mean excess. A candidate is considered as the threshold when the line appears to be unstable. It is hard to apply this interpretation in Figure 3 because the unstableness of the mean excess starts to begin around u=200. After u=200, MRLP appears more unstable. We can estimate that the threshold could be between 150 and 200, meanwhile the true threshold is at 100. The reason why this bias estimation happens is because the bulk distribution (Truncated Gamma) and the GPD has a very subtle change. This means that the MRLP is unable to precisely estimate the exact border and can only notice the change after a clear difference at u=200.

4.2 Rainfall Case

The true threshold for the rainfall data is unknown. We first started with an exploration analysis. Figure 4 shows the histogram and density plot of rainfall to explore the probable bulk and upper distribution. In general, rainfall data is not categorized as a normal distribution, so the other option is Gamma distribution. Stephenson and his colleagues [15] state that the Gamma and Weibull distribution provide a good fit to the rainfall distribution. The bulk distribution in Figure 4 looks closer to Gamma distribution. After the largest peak at 80-90 mm, several smaller peaks were found in the distribution. This condition makes it more difficult to determine the threshold by exploration. There were some points where the distribution started to climb again around 140 mm and 210 mm. This will make the interpretation of the measure of surprise diagram more challenging.

Measure of surprise diagram for rainfall data is in Figure 5. The first 7 of the 29 candidates (0 – 60) shows a small decrease of surprise from 0.01 (u=0) to 0.2 (u=60) which means that the distribution is not yet GPD at this

interval. There is a 0.11 surprise change between 60 and 70. Candidate 70 until 110 is relatively stable at p=0.3. Candidate 140 hits p=0.53 which indicates that at this point, the distribution might start to be GPD. However, the surprise after 140 is not stable around 0.5 where candidates 150-190 is fluctuating in the range of 0.35 and 0.43. This is parallel with the trend in the histogram and density plot. Surprise at candidate 210 is 0.5 and the average of 210 until 280 is 0.51 ± 0.02. This shows that the surprise after 210 is stable and therefore concludes that the GPD starts from this point. The candidates are only until 280 as the number data observed above 280 is only 25 which is too small.

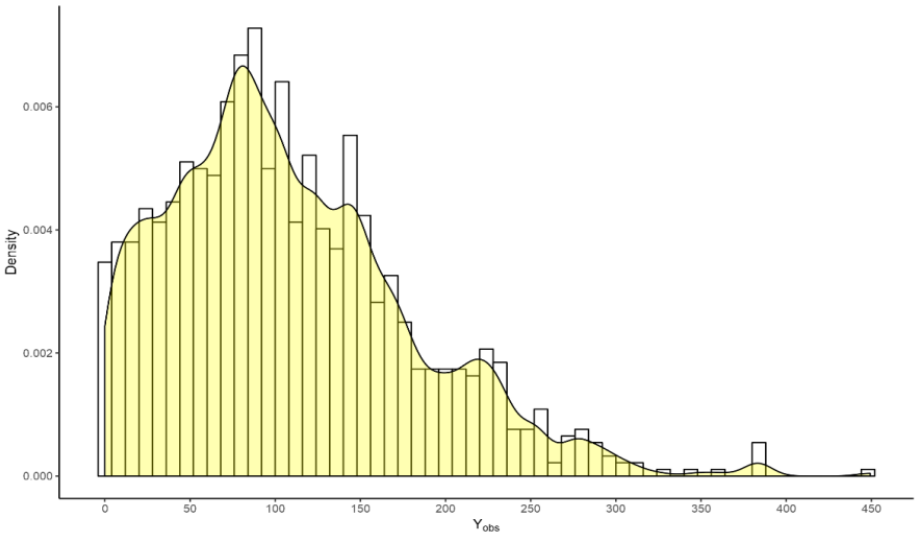

Figure 4: Histogram and Density Plot of Rainfall Data for the MCMC to run. The total observation above 210 is 118 observation.

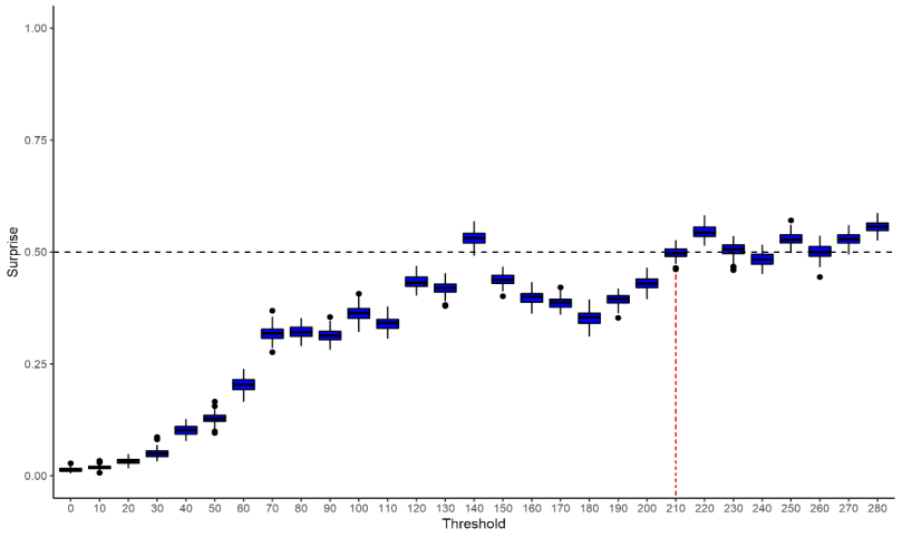

Figure 5: Rainfall Measure of Surprise Diagram

Finally, we compare measure of surprise of rainfall data to MRLP. The interpretation of MRLP in Figure 6 appears to be ambiguous. We find it hard to determine when the mean exceed starts to become non-linear. There are three points between 100 and 200 which shows signs of instability, but we still find some stability between

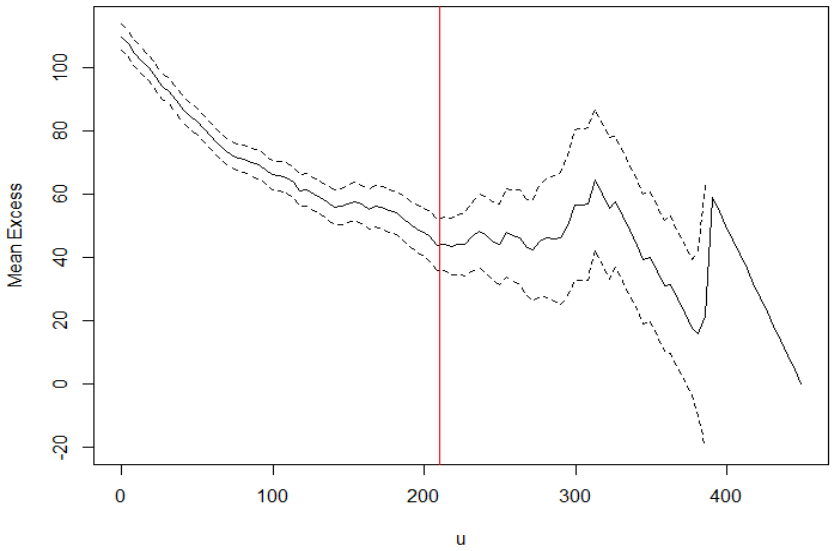

Figure 6: Mean Residual Life Plot of Rainfall Data

these three points. The graph shows a clear instability on some points beyond 210 but it is still hard to decide the exact point where instability begins. The red line at u=210 helps to see the threshold estimation from the measure of surprise. We might tend to overestimate the threshold around 230 or above if we merely just use the MRLP method.

5. Conclusion

Measure of surprise is a useful method to estimate the GPD threshold as it can be easily interpreted. The estimation process is fully Bayesian. Threshold estimation for simulation data shows an appropriate result between the estimation and the real quantity. The estimation for rainfall data shows that the extreme rainfall starts from 210 mm. The measure of surprise method is easier to interpret than MRLP which lead to ambiguity in determining the point of instability.

Acknowledgment

The writer would like to thank Jeong E. (Kate) Lee and Scott Sisson for the discussion regarding the measure of surprise and R code for calculation measure of surprise with Bayesian Method and MCMC Metropolis-Hastings algorithm.

References

[1] S. Coles. An Introduction to Statistical Modeling of Extreme Values. London, UK: Springer, 2001.

[2] I. Vicari. “Measure of surprise and threshold selection in extreme value statistics”. Thesis. École Polytechnique Fédérale de Lausanne, 2010.

[3] J. Pickands. “Statistical inference using extreme order statistics”. Annals of Statistics, vol. 3, pp. 119-131, 1975.

[4] A.C. Davison, R.L. Smith. “Models for exceedances over high thresholds”. Journal of the Royal Statistical Society, Series B, vol. 52, pp 393-442, 1990.

[5] C.N. Behrens, F.L. Hedibert, and D. Gamerman. “Bayesian analysis of extreme events with threshold estimation”. Statistical Modelling, vol. 4, pp. 227-244, 2004.

[6] C.J. Scarrott, A. MacDonald. “A review of extreme value threshold estimation and uncertainty quantification”. Statistical Journal, vol. 10, pp. 33-60, 2012.

[7] W.H. DuMouchel. “Estimating the stable index α in order to measure tail thickness: A Critique”. The Anns of Statistics, vol.11, pp.1019-1031, 1983

[8] W. Weaver. “Probability, rarity, interest and surprise”. Scientific Monthly. vol. 67, pp. 390-392, 1948.

[9] G.E.P Box. “Sampling and bayes inference in scientific modeling and robustness”. Journal of the Royal Statistical Society, Series A, vol. 143, pp. 383-430, 1980.

[10] M. J. Bayarri, J.O.Berger. “Measure of Surprise in Bayesian Analysis”. ISDS Discussion Paper, 1997, pp. 97-154.

[11] X. L. Meng. “Posterior predictive p-values”. The Annals of Statistics, vol. 22, pp. 1142-1160, 1994.

[12] J. Lee, Y. Fan, S.A. Sisson. “Bayesian threshold selection for extremal models using measures of surprise”. Computational Statistics and Data Analysis, vol. 85, pp. 84-99, 2015.

[13] I. Guttman. “The use of the concept of a future observation in goodness-of- fit problems”. Journal of the Royal Statistical Society, Series B, vol. 29, pp. 83-100, 1967.

[14] M. Castellanos S. Cabras, J. “A default Bayesian procedure for the generalized Pareto distribution”. Journal of Statistical Planning and Inference, vol. 137, pp. 473-483, 2007.

[15] D.B. Stephenson, K.R. Kumar, F.J. Doblas-Reyes, J.F. Royer, E. Chauvin, S. Pezzulli. “Extreme Daily Rainfall Events and Their Impact on Ensemble Forecasts of the Indian Monsoon”. Monthy Weather Review, vol. 127, pp. 1955-1966, 1999.